Journal of Database Management
Volume 36 • Issue 1 • January-December 2025

FADE:
Focused and Attention-Based Detector of Errors

Omar Al-Shamali
University of Alberta, Canada

James Miller
University of Alberta, Canada

Shaikh Quader
IBM Canada, Canada

ABSTRACT

Error detection is an important part of preparing data for data analysis. Erroneous data can result in inaccurate analysis, resulting in garbage-in, garbage-out. Currently, many models utilize either or both Qualitative and Quantitative methods to detect errors in the data. However, these methods are still limited in the errors they can detect. Hence FADE, Focused and Attention-based Detector of Errors, was proposed. FADE can detect errors within structured data with rows and columns. FADE utilizes the information found in surrounding cells within the same row to help determine if a cell is erroneous. It also learns the expected structure of the attributes in the dataset and the values expected in each attribute. This results in FADE having a much wider range of error type detection and having a higher classification of errors than other methods. FADE was evaluated and was found to detect these errors with relatively high performance.

KEYWORDS

Data Cleaning, Error Detection, Tabular Data, AI, Artificial Intelligence, Machine Learning, Neural Networks, Data Augmentation

INTRODUCTION

We currently live in the information age, where we handle and process huge and diverse amounts of data. Examples of data sources include the internet, social media applications, the Internet of Things, databases, and many more. It is crucial to perform a proper analysis of such "big" data in a manner that is timely and accurate since that data is also increasingly helping us make important decisions, such as medical, loan, and hiring decisions (Restat et al., 2022). Sixty-two percent of marketing decision-makers agree that the most important factor for a successful marketing data strategy is improving the quality of the data for marketing (Crane, 2017). However, low-quality inputted data can reduce the viability of outputted information impairing any analysis. This can cause great financial loss (Eckerson, 2002) and can have a serious negative impact on the world of business (Redyuk et al., 2021). The result is garbage-in, garbage-out, which restricts the analytical power of software systems. Therefore, it is crucial to detect errors in the data before using it. This raises the importance of finding effective techniques to enhance the quality of data.

There are numerous works in the literature, which employ different methodologies for detecting faulty, missing, and misplaced data within a database or structured data sources. One category of

DOI: 10.4018/JDM.377526

systems uses qualitative methods, which utilize rule-based techniques in data analysis, such as in the works of Oliveira et al. (2020), Pena et al. (2020), Fan et al. (2021), Qahtan et al. (2020), Zheng et al. (2021), and Abu Ahmad and Wang (2020). However, such systems are limited to detecting errors that violate the rules defined by a user, and the performance is dependent upon these user-defined rules. Another category includes systems that employ quantitative methods, which utilize statistical techniques to detect erroneous data, such as in the works of Redyuk et al. (2021) and Sun et al. (2020). Such methodologies, however, require errors to exist in relatively small amounts within the data to ensure that the model detects the errors as outliers. There are also hybrid methods, which utilize both qualitative and quantitative techniques, such as in the works of Ge et al. (2020) and Yan et al. (2020). However, these systems still have the same limitations as in the standalone qualitative and quantitative categories.

Our research project builds on previous work, which introduced TabReformer. TabReformer is a system that utilizes bidirectional encoder representations from transformers (BERT; Devlin et al., 2019), which is a neural network model used in natural language processing tasks. BERT was modified and repurposed to become TabReformer, whose function is to detect erroneous cells in tabular datasets. An important characteristic of BERT is that it utilizes the attention mechanism (Vaswani et al., 2017) to find relationships between words. This feature is what allows TabReformer to recognize error patterns within cell values and classify errors in a dataset. TabReformer starts by taking each cell as input and then applies the attention mechanisms to deduce whether it is erroneous or not. However, there are types of error that cannot be identified just by looking at the target cell alone. The assessment of a faulty cell often requires checking the surrounding cells. Examples of these error types that require the information from surrounding cells include value swapped across columns (VS), values violating data constraints (VD), formatting errors (FE), and missing values (MV).

In our work, we introduce a new error detection model called focused and attention-based detector of errors (FADE). This new model presents two significant contributions to error detection:

- FADE has the unique ability to utilize the information in a specific target cell and its surrounding cells during the error detection process. More specifically, FADE takes, as input, the target cell being classified and all the other cells within the same tuple (same row). Next, our model uses the attention mechanism to assess the information and context collected from these cells and then classifies the target cell as correct or erroneous. Feeding the model the whole tuple is an important contribution of our work. It widens the scope of our model by increasing its ability to recognize erroneous cells. This gives our model the ability to detect a greater range of error types that are out of reach for the older TabReformer, which only looks at the target cell.
- Additionally, we created a focus feature that allows our model to classify the cells, one cell at a time, enhancing classification accuracy. Using this feature, our model can classify a single cell, even when a whole tuple is given as input. Such a protocol has the advantage of reducing the size of the model's output by keeping intermediate processed information in binary format. This improves its scaling power to include large data without requiring to increase in the size of the model's output layer.

In this report, the Related Work section is dedicated to researching the literature and reviewing related work. In the FADE: The Complete System section, we described our proposed model in detail and discussed how it evolved from previous work. Methodology section provides a detailed description of the experiments we conducted using our model. This is followed by the Results and Discussion section, which states the experimental results we achieved, and discussions about their significance. The limitations of the model are discussed in the Threats to Validity section, and, finally, the Conclusions section is dedicated to the conclusion.